شناسایی جوامع در گراف‌های ارتباطی با استفاده از الگوریتم IsoFdp بهبود یافته

رضا شمسایی [*۱]، ماهنوش خوشخو[۲].

۱- عضو هیئت علمی، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی سجاد، مشهد

۲- دانشجوی کارشناسی ارشد مهندسی کامپیوتر نرم‌افزار، دانشگاه صنعتی سجاد، مشهد

چکیده

در سالهای گذشته، شبکه‌های تعاملی با توجه به یال‌های ارتباطی که میان گره‌های آن وجود دارد مورد توجه زیادی قرار گرفته است. تشخیص جامعه یکی از چالش‌های اصلی برای تجزیه و تحلیل اینگونه شبکه‌ها است. جامعه زیرگرافی از یک گراف است که تعداد ارتباط‌های بین اعضاء آن زیرگراف، نسبت به تعداد ارتباط‌هایی که آن را به بقیه گراف متصل می‌کند، بیشتر است. هدف اکثر الگوریتم‌های تشخیص جامعه، تقسیم‌بندی گراف به چند زیرگراف همبند است که هر گره باید به یک جامعه متعلق بوده و به دیگر جوامع تعلق نداشته باشد. جامعه‌بندی گره‌ها بر اساس شباهت‌های موجود در بین آنها انجام می‌شود. کلید اصلی جامعه‌بندی صحیح، استفاده از معیار شباهت مناسب برای تفکیک گره‌ها است. در این مقاله از یک معیار شباهت ترکیبی برای جامعه‌بندی گره‌ها استفاده شده است. این معیار بر اساس ترکیب وزن‌دار معیار شباهت‌های ساختاری، کسینوسی و همینگ ایجاد شده است. نتایج مقایسات نشان‌دهنده عملکرد بهتر روش پیشنهادی در مقایسه با موارد مشابه است.

کلمات کلیدی: تشخیص جامعه، گراف ارتباطی، الگوریتم IsoFdp، معیار شباهت ترکیبی.

۱. مقدمه

بسیاری از مجموعه داده‌هایی که از حوزه‌های مختلف جهان واقعی نشات می‌گیرند، می‌توانند به صورت شبکه‌های ارتباطی (تعاملی) ظاهر شوند. یک شبکه ارتباطی ساختاری است که موجودیت‌های درون آن با یکدیگر ارتباط داشته، از جمله اعضای یک حزب سیاسی که دارای ارتباطات درون حزبی می‌باشند [۱].

با رشد شبکه‌های اجتماعی، کاربران زیادی به اینگونه شبکه‌ها جذب شده‌اند. این شبکه‌ها محل تبادل اطلاعات بسیار زیادی بین افراد هستند. با توجه به این نکته که رفتار و تعاملات افراد تنها بر اساس ویژگی‌های شخصی آنان نیست (بلکه تعاملات آنها با دیگران بر رفتارشان تاثیرگذار است)، تجزیه و تحلیل شبکه‌های ارتباطی بسیار مورد توجه قرار گرفته است. شبکه‌هایی مانند فیسبوک، توییتر و لینکداین از جمله شبکه‌های ارتباطی بزرگ هستند [۲].

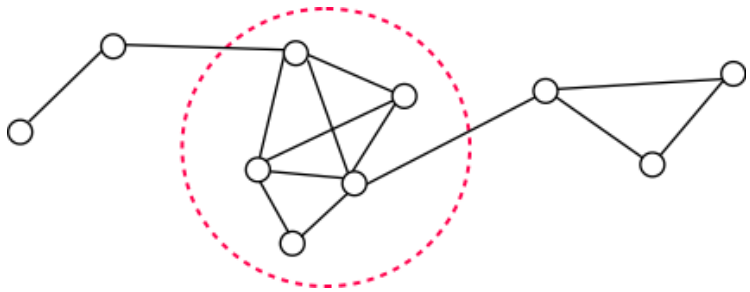
* **Email: r_shamsaee@sadjad.ac.ir**

یکی از مسائل بسیار مهم در تجزیه و تحلیل شبکه‌های ارتباطی، شناسایی ساختارهای موجود در آن است. واضح‌ترین ساختار درون شبکه ارتباطی، ساختار جامعه است، در نتیجه مسئله تشخیص جامعه، یکی از معمول‌ترین تحقیقات در شبکه‌ها است. تشخیص و شناسایی جوامع موجود، به درک ساختار کلی شبکه و تقسیم‌بندی فعالیت‌هایی که در آن صورت می‌پذیرند، می‌انجامد [۳].

ساده‌ترین روش مطالعه شبکه‌های ارتباطی، نگاشت آنها بر تئوری گراف در ریاضیات است. از دیدگاه ریاضی، یک گراف G که با نماد $G(V, E)$ نشان داده می‌شود، از دو مجموعه مجزا به نام‌های، مجموعه گره‌ها (V) و مجموعه یال‌ها ($E \subset V \times V$) تشکیل شده است. هر یال از مجموعه E، دو گره از گراف را به یکدیگر وصل می‌کند [۴]. در این نگاشت، موجودیت‌های شبکه (به‌طور مثال افراد) به‌صورت گره‌های گراف و روابط بین موجودیت‌ها به‌صورت یال‌های گراف تعریف می‌شوند. روابط می‌تواند یک‌طرفه یا دوطرفه باشد. روابط یک‌طرفه به یال‌های جهت‌دار و روابط دوطرفه به یال‌های بدون‌جهت در گراف نگاشت می‌شوند. با توجه به نگاشت شبکه به یک گراف، مسئله تشخیص جامعه در واقع خوشه‌بندی ارتباطات یک گراف است [۵].

از نگاه بیرون به مسئله، جامعه‌بندی مانند خوشه‌بندی است. اما در نگاه دقیق‌تر جامعه‌بندی یک نوع خوشه‌بندی خاص است (خوشه‌بندی گراف بر اساس یال)، و با خوشه‌بندی محض متفاوت است. در مفهوم خوشه‌بندی، نمونه‌های داده بر اساس شباهت بین ویژگی‌های خود داده، خوشه‌بندی می‌شوند (خوشه‌بندی گراف بر اساس ویژگی‌های گره). اما جامعه‌بندی بر اساس ارتباطات گره‌ها (یال‌ها) انجام می‌شود و معیار شباهت، وابسته به همین ارتباطات است [۶].

تاکنون هیچ تعریفی برای مفهوم جامعه به صورت عمومی مورد پذیرش قرار نگرفته است. اما به طور کلی جامعه زیرگرافی از یک گراف است که تعداد یال‌های بین اعضاء در زیر گراف، نسبت به تعداد یال‌هایی که زیرگراف را به بقیه گراف متصل می‌کند، بیشتر است (با فرض همبند بودن زیرگراف) [۷]. از دیدگاه ریاضی، گراف $H(V1, E1)$ را زیرگراف $G(V, E)$ است اگر و فقط اگر $V_1 \subset V$ و $E_1 \subset E$ آنگاه $H \subset G$ است [۸]. در شکل ۱ یک زیرگراف به‌صورت خط‌چین از بقیه گراف متمایز شده است و می‌تواند یک جامعه باشد.



شکل ۱- دایره خط‌چین زیرگرافی از گراف اصلی را نشان می‌دهد که معرف یک جامعه است.

هدف جامعه‌بندی، گردآوری گره‌ها در گروه‌های مشخص و مجزّا به نام جامعه است. زیرگراف‌های تشخیص داده‌شده بیانگر جامعه‌های موجود در گراف هستند. دسته‌بندی نمونه داده‌ها در جوامع موجود، بر اساس شباهت‌های موجود در بین آنها انجام می‌شود. با توجه به این رویکرد دسته‌بندی، هر جامعه شامل مجموعه‌ای از گره‌ها است که وجه اشتراک بیشتری با یکدیگر، نسبت به بقیه گره‌ها دارند. کاربردهای جامعه‌بندی در فعالیت‌های مانند سیستم‌های پیشنهاددهنده، بهینه‌سازی موتورهای جستجو و پیش‌بینی رفتار گروهی جوامع می‌باشد.

یکی از مسائل مهم در تشخیص زیرگراف‌ها، تعیین میزان شباهت بین گره‌ها است. اگر شباهت بین گره‌های یک گراف به‌درستی اندازه‌گیری شود موجب بهبود دقت و کیفیت جامعه‌بندی خواهد شد. معیارهای شباهت متفاوتی برای این کار ارائه شده‌اند. اساس معیار شباهت در جامعه‌بندی، ارتباطات هر گره با گره‌های دیگر است [۹].

در این مقاله روشی برای جامعه‌بندی گراف‌های بدون جهت، بر اساس الگوریتم IsoFdp ارائه شده است. در این روش با تغییر معیار شباهت، از یک معیار ترکیبی برای جامعه‌بندی گره‌ها استفاده شده است. این معیار بر اساس ترکیب وزن‌دار معیار شباهت‌های ساختاری، کسینوسی و همینگ ایجاد شده است.

ساختار ادامه مقاله به این شرح است: بخش دوم مروری بر کارهای گذشته این حوزه است. روش پیشنهادی در بخش سوم توضیح داده می‌شود. بخش چهارم ارزیابی روش پیشنهادی است و نتیجه‌گیری در بخش پنجم ذکر شده است.

۲. مروری بر ادبیات پیشین

یک دسته‌بندی کلی روش‌های تشخیص جامعه بر اساس نوع گراف (جهت‌دار و بدون‌جهت) است. در این مقاله، اساس کار بر پایه گراف‌های بدون جهت است.

الگوریتم گیروان و نیومن که در [۱۰] معرفی شده است، یکی از روش‌های خوشه‌بندی مورد استفاده برای شناسایی جوامع است. بینابینی گره، به عنوان یک معیار مرکزیت و قابلیت نفوذ گره‌ها مورد مطالعه قرار گرفته است. این معیار تاثیر گره بر جریان اطلاعات بین گره‌ها، به خصوص در مواردی که جریان اطلاعات عمدتا در کوتاهترین مسیر در دسترس است را بیان می‌کند. الگوریتم گیروان و نیومن این تعریف را به یال گسترش داد. بینابینی یک یال به عنوان تعداد دفعاتی که آن یال در کوتاهترین مسیر بین هر جفت از گره‌ها حادث می‌شود، تعریف می‌شود. اگر بیش از یک کوتاهترین مسیر بین یک جفت از گره‌ها وجود داشته باشد، به هر یک از مسیرها یک وزن مساوی اختصاص داده می‌شود، به طوری که وزن کل تمام مسیرها برابر واحد است. در این الگوریتم بیان شده که یال‌های انسانی جوامع، بینابینی یال، بالایی (حداقل یکی از آنها) دارند. با از بین بردن این یال‌ها، گروه‌ها از یکدیگر جداشده و بنابراین ساختار جامعه در شبکه شکل می‌گیرد. نتیجه نهایی الگوریتم گیروان و نیومن یک دندوگرام[۱] است. همانطور که الگوریتم گیروان و نیومن اجرا می‌شود، دندوگرام از بالا به پایین تولید می‌شود. برگ‌ها در دندوگرام گره‌های فردی هستند.

در مقاله [۱۱]، یک روش خوشه‌بندی جدید برای تشخیص ساختار جامعه در شبکه معرفی می‌شود که استفاده از آن برای تجزیه و تحلیل برخی از شبکه‌های اجتماعی، مناسب است. در این روش، افراد و روابط آنها توسط گراف وزن‌دار نشان داده شده است. این روش در مقایسه با روش‌های موجود، دارای دو ویژگی برجسته است:

- ایجاد درختان سلسله مراتبی بسیار کوچکتر که به وضوح خوشه‌های معنی‌دار را نمایش می‌دهند.
- تامین هم‌پوشانی خوشه‌ها که بازتابی از پیچیدگی‌های دنیای واقعی است.

در مقاله [۱۲] یک استراتژی برای بهبود الگوریتم‌های تشخیص جامعه موجود را با اضافه‌کردن یک مرحله پیش‌پردازش که در آن به یال‌ها با توجه به مرکزیتشان در توپولوژی شبکه وزن داده می‌شود، پیشنهاد شده است. در این روش، مرکزیت یال نشان‌دهنده همکاری آن در تراگذاری گراف است. محاسبه مرکزیت یال با انجام پیاده‌روی‌های متعدد تصادفی با طول محدود بر روی شبکه انجام می‌شود که باعث می‌شود، محاسبات مرکزیت یال در شبکه‌هایی در مقیاس بزرگ نیز عملی شود.

در [۱۳] روشی به‌نام Fdp ارائه شده است. این روش بر اساس DBSCAN کار می‌کند. روش DBSCAN نقاط را به سه دسته نقاط هسته، نقاط مرز و نقاط نویز تقسیم می‌کند. تقسیم نقاط توسط پارامترهای Eps و MinPts انجام می‌شود. Eps شعاع همسایگی هر نقطه را در مجموعه داده مشخص می‌کند. نقاطی که در این شعاع قرار می‌گیرند، نقاط همسایه خواهند بود. نقاطی که بیش از MinPts همسایه دارند به عنوان نقاط هسته، شناخته می‌شوند. نقاطی که هسته

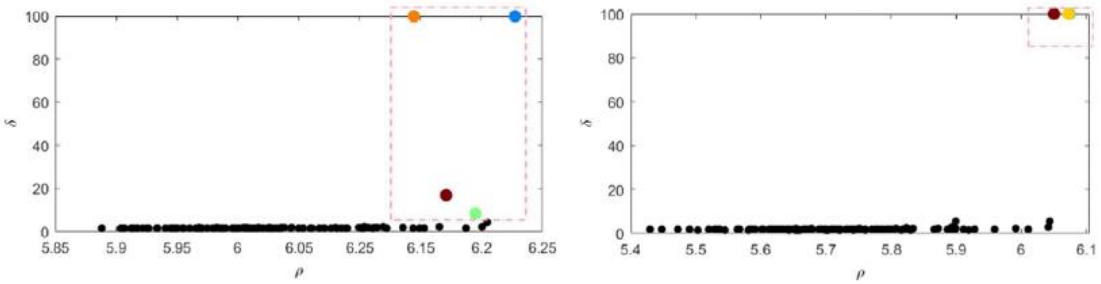
نیستند اما در همسایگی یکی از نقاط هسته قرار دارند به عنوان نقاط مرز شناخته می‌شوند و نهایتا نقاطی که هسته و مرز نیستند نقاط نویز خواهند بود.

در روش Fdp از دو پارامتر ρ_i و δ_i برای تشخیص جوامع استفاده می‌کند. این پارامترها برای هر یک از گره‌های شبکه توسط رابطه‌های (۱) و (۲) تعیین می‌شوند.

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \quad (۱)$$

$$\delta_i = \min(d_{ij}),\ j : \rho_i < \rho_j \quad (۲)$$

در رابطه (۱)، d_c شعاع همسایگی و d_{ij} فاصله بین گره i و j است. زمانی که $d_{ij} - d_c < 0$ باشد، $\chi(d_{ij} - d_c)$ برابر ۱ خواهد بود، در غیر اینصورت صفر خواهد بود. ρ_i در واقع نشان دهنده چگالی محلی داده i است و δ_i کوتاهترین فاصله میان نقطه i با نقاطی است که دارای چگالی محلی بالاتری هستند. شکل ۲ نمایی از مقادیر ρ_i و δ_i را برای نقاط مختلف نشان می‌دهد. این نمودار به گراف تصمیم گیری معروف است.



شکل۲- گراف تصمیم‌گیری براساس مقادیر ρ_i و δ_i [۱۴]

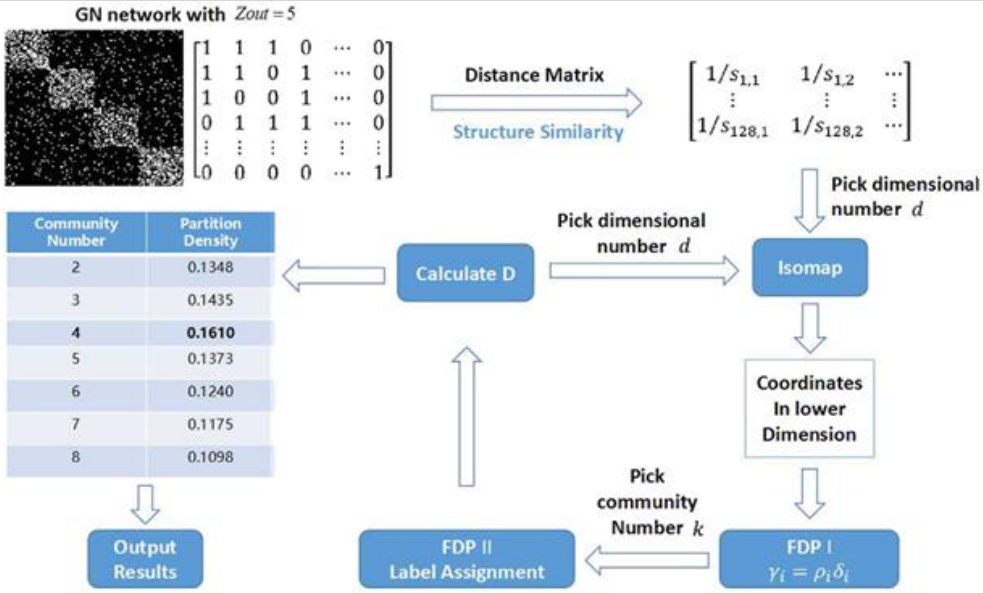
مقادیر ρ_i و δ_i برای مراکز خوشه‌ها، به میزان قابل توجهی بیشتر از دیگر نقاط است. از اینرو، کاربران می‌توانند با نگاه به گوشه بالا سمت راست گراف تصمیم‌گیری، مراکز خوشه‌ها را تشخیص دهند. با این روش می‌توان پس از تعیین مراکز خوشه‌ها نقاط باقیمانده را با توجه به فاصله از این مراکز به خوشه‌ها تخصیص داد.

زمانی که مجموعه داده، حاوی ابعاد بالایی باشد و گره‌های مشابه در این مجموعه داده زیاد باشند روش Fdp عملکردی مطلوب از خود نشان نمی‌دهد. عملکرد نامطلوب روش Fdp به این علت است که، در این حالت فاصله معنایی میان گره‌ها بسیار کم می‌شود. از اینرو، در مقاله [۱۴] به‌نام روش IsoFdp، از رابطه (۳) برای تعیین شباهت میان گره‌ها استفاده می کند.

$$SS(v,w) = \frac{|N(v) \cap N(w)|}{\sqrt{|N(v) \times N(w)|}} \quad ,N(v) = \{w \in V \mid (v,w) \in E\} \cup \{v\} \quad (۳)$$

$N(v)$ شامل گره‌هایی است که با گره v ارتباط دارند به همراه خود گره v. $SS(v,w)$ نشان‌دهنده شباهت میان گره‌های v و w است (شباهت ساختاری). $N(v) \cap N(w)$ تعداد گره‌هایی است که با هر دو گره v و w در ارتباط هستند. روال کاری روش IsoFdp در شکل ۳ مشخص شده است. در ادامه این روش بررسی خواهد شد.

روش IsoFdp با دریافت مجموعه داده در مرحله اول میزان شباهت هر دو گره در مجموعه داده را محاسبه می‌کند. با محاسبه میزان شباهت میان گره‌های v و w، IsoFdp از رابطه $\frac{1}{SS(v,w)}$ برای محاسبه ماتریس فاصله میان هر دو گره استفاده می‌شود.



شکل ۳- نمایی کلی از روش IsoFdp [۱۴]

در ادامه IsoFdp وارد مرحله Isomap می‌شود. در این مرحله، ابتدا با استفاده از الگوریتم کوتاهترین فاصله فلوید و ماتریس فاصله DG، کوتاهترین فاصله میان هر دو گره را محاسبه می‌کند. نتیجه این محاسبات در ماتریس DG` ذخیره خواهد شد. در الگوریتم فلوید زمانی که میان دو گره مسیری کوتاهتر به واسطه گره سومی ایجاد شود مسیرها اصلاح می‌شود، به نحوی که برای انتقال از گره اول به گره دوم از گره سوم به عنوان گره واسطه استفاده شود.

پس از یافتن کوتاهترین مسیر میان گره‌ها، در مرحله Isomap با استفاده از الگوریتم MDS نتایج بدست آمده در DG`، به مختصات نقاط تبدیل می‌شود. روش MDS، جهت تبدیل فاصله میان مجموعه گره‌ها به مختصات گره‌ها استفاده می‌شود. در واقع، عملکرد روش MDS در مقابل روش‌های تعیین فاصله قرار می‌گیرد. زیرا در این روش، ماتریسی حاوی فاصله میان گره‌ها دریافت می‌شود و خروجی، مختصات گره‌ها در d بعد است. پس از مرحله Isomap روش IsoFdp، از روش Fdp دو مرحله‌ای استفاده می‌کند.

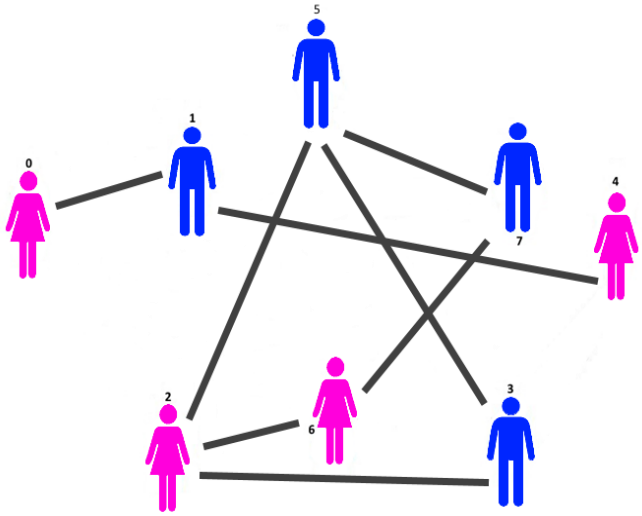
در مرحله اول از Fdp دومرحله‌ای، در عوض استفاده از پارامترهای ρ_i و δ_i به صورت جداگانه از حاصل‌ضرب آنها (γ_i) برای تعیین شرایط تقسیم گره‌ها به گره‌های هسته، مرز و نویز استفاده می‌شود. پس از آن با استفاده از γ_i و روش DBSCAN گره‌ها خوشه‌بندی می‌شوند. روش IsoFdp برای محاسبه میزان مطلوبیت جامعه‌بندی ایجاد شده، از پارامتر چگالی که توسط رابطه (۴) تعیین می‌شود، استفاده می‌کند.

$$D = \frac{2}{\sqrt{k}\,N} \sum_{c} m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \qquad (۴)$$

در این رابطه c نشان دهنده جوامع ایجاد شده است، k تعداد جوامع و N تعداد گره‌هاست. n_c و m_c به ترتیب نشان‌دهنده تعداد گره‌ها و یال‌ها در جامعه c است. همانطور که در شکل ۳ مشخص است روش IsoFdp تلاش می‌کند تا به ازای تعداد جوامع مختلف عمل جامعه‌بندی را انجام دهد و نتایج را بر اساس مقدار چگالی بدست آمده با یکدیگر مقایسه کند.

۳. روش پیشنهادی

جامعه را می‌توان مجموعه‌ای از اجزاء، به همراه روابط میان آنها تعریف کرد. از اینرو، هر جامعه‌ای از دو قسمت اصلی تشکیل شده است: اجزاء و روابط میان آنها. هر جزء جامعه، یک گره و هر ارتباط میان دو گره، توسط یک یال نمایش داده می‌شود [۱۵]. برای مثال جامعه‌ای متشکل از ۸ فرد با روابط میان آنها در شکل ۴ نشان داده شده است.



شکل۴- یک جامعه فرضی با ۸ گره و ارتباطات میان آنها

روش پیشنهادی این مقاله، در واقع بهبود یافته روش IsoFdp است. روش پیشنهادی تلاش دارد تا گام دوم روش IsoFdp یعنی محاسبه معیار شباهت را بهبود ببخشد. همانطور که نتایج ارائه شده در روش IsoFdp [۱۴] نشان می‌دهد، در مقایسه میان روش‌های محاسبه شباهت، روش محاسبه شباهت Structure (که به شباهت SSIM معروف است) بهترین نتایج را ارائه می‌دهد. پس از آن شباهت کسینوسی نتایج بهتری را نسبت به دیگر روش‌های محاسبه شباهت داراست. نهایتاً شباهت همینگ در مقام سوم قرار می‌گیرد. از اینرو، در این مقاله به منظور بهبود روش IsoFdp از ترکیب وزن‌دار شباهت‌های SSIM، کسینوسی و همینگ استفاده می‌شود. در روش پیشنهادی از رابطه (۵) به عنوان رابطه محاسبه شباهت، جایگزین شباهت به کار رفته در روش IsoFdp استفاده خواهد شد.

$$S = \alpha \times SSIM + \beta \times Cos + \gamma \times Ham \qquad (۵)$$

در رابطه (۵) SSIM شباهت ساختاری، Cos شباهت کسینوسی و Ham شباهت همینگ است. مجموع ضرایب α, β, γ در رابطه (۵) باید مقدار یک شود. در نتیجه با تغییر ضرایب می‌توان اهمیت و تاثیر شباهت‌های استفاده شده را تغییر داد. با تغییر مقادیر ضرایب می‌توان مناسب‌ترین حالت ترکیب سه معیار شباهت را برای رسیدن به یک معیار شباهت کاراتر در روش پیشنهادی محاسبه کرد.

در ابتدا ماتریس مجاورت میان گره‌ها بر اساس ارتباط میان گره‌ها ایجاد می‌شود. برای مثال شبکه ارتباطی شکل ۴ مفروض است. ماتریس مجاورت این شبکه ارتباطی همانند آنچه در جدول ۱ نشان داده شده است خواهد بود. پس از محاسبه ماتریس مجاورت میان گره‌ها، شباهت‌های SSIM، کسینوسی و همینگ محاسبه می‌شوند.

جدول ۱- ماتریس مجاورت گره‌های شبکه شکل۴

	۰	۱	۲	۳	۴	۵	۶	۷
۰	۰	۱	۰	۰	۰	۰	۰	۰
۱	۱	۰	۰	۰	۱	۰	۰	۰
۲	۰	۰	۰	۱	۰	۱	۱	۰
۳	۰	۰	۱	۰	۰	۱	۰	۰
۴	۰	۱	۰	۰	۰	۰	۰	۰
۵	۰	۰	۱	۱	۰	۰	۰	۱
۶	۰	۰	۱	۰	۰	۰	۰	۱
۷	۰	۰	۰	۰	۰	۱	۱	۰

۱،۳. شباهت ساختاری

همانطور که از نام شباهت ساختاری مشخص است از آن برای تعیین میزان شباهت ساختاری میان دو گره استفاده می‌شود. محاسبه شباهت ساختاری توسط رابطه (۳) محاسبه می‌شود. برای مثال شباهت ساختاری میان گره‌های ۲ و ۷ در ماتریس مجاورت جدول ۱ به صورت زیر محاسبه می‌شود.

$$N(2)=\{2,3,5,6\}$$
$$N(7)=\{5,6,7\}$$
$$N(2)\bigcap N(7)=\{5,6\}$$
$$SS(2,7)=\frac{2}{\sqrt{4\times 3}}=0.578$$

۲،۳. شباهت کسینوسی

از این روش برای محاسبه میزان شباهت میان دو بردار غیر صفر استفاده می‌شود. برای این منظور از رابطه (۶) استفاده می‌شود.

$$Cosine\ Similarity\ (v,w)=\frac{\sum_{i=1}^{n}A_{vi}\ A_{wi}}{\sqrt{\sum_{i=1}^{n}A_{vi}^{2}}\ \sqrt{\sum_{i=1}^{n}A_{wi}^{2}}} \qquad (۶)$$

در این رابطه A_{vi} نشان دهنده مقدار متناظر با سطر v و ستون i در ماتریس مجاورت است و n تعداد گره‌های شبکه را نشان می‌دهد. ذکر این نکته ضروری است که بردار ارتباط برای هر گره، معادل با سطر آن گره در ماتریس ارتباط است. از اینرو با استفاده از بردار ارتباط هر دو گره می‌توان شباهت میان هر دو گره را محاسبه کرد.

۳،۳. شباهت همینگ

شباهت همینگ براساس میزان شباهت میان ساختار دو رشته محاسبه می‌شود. برای محاسبه شباهت همینگ برای دو رشته با طول مساوی، ابتدا تعداد مکانهای متناظر که دارای مقادیر مشابه هستند، محاسبه می‌شود سپس این مقدار تقسیم

بر طول رشته می‌شود. هر سطر از ماتریس ارتباط نشان دهنده یک رشته صفر و یک است. این رشته در واقع میزان ارتباط میان گره‌ها را نشان می‌دهد در نتیجه می‌توان میزان شباهت همینگ میان هر دو گره را محاسبه کرد.

بعد از محاسبه شباهت، الگوریتم IsoFdp اجرا می‌شود. با فراخوانی IsoMap و اجرای الگوریتم دومرحله‌ای از Fdp جامعه‌بندی گره‌ها انجام می‌شود.

۴. ارزیابی و مقایسه روش پیشنهادی

در این بخش روش پیشنهادی با روش IsoFdp مورد مقایسه قرار گرفته است. برای ارزیابی روش پیشنهادی از معیار چگالی جوامع، اطلاعات متقابل نرمال شده (NMI) و صحت، جهت تشخیص مطلوبیت جامعه‌بندی استفاده می‌شود. چگالی، یک شاخص کیفیت برای تشخیص اجتماع است. در جامعه‌بندی مطلوب، بایستی تلاش شود تا ارتباطات میان گره‌های جامعه بیشتر (جامعه چگال‌تر) و میان گره‌های بین جوامع، کم باشد. معیار چگالی تلاش می‌کند این ویژگی را در جامعه بندی‌های ارائه شده نشان دهد. رابطه (۴) نحوه محاسبه معیار چگالی را نشان می‌دهد [۱۶].

NMI که توسط رابطه (۷) محاسبه می‌شود، میزان نزدیکی جامعه‌بندی نهایی حاصل از سیستم پیشنهادی با جامعه‌بندی بهینه را مشخص می‌کند [۱۷].

$$NMI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} \ln\left(\frac{n_{ij}.n}{n_i^c.n_j^{c'}}\right)}{\sqrt{\left(\sum_{i=1}^{k} n_i^c \ln(n_i^c/n)\right)\left(\sum_{j=1}^{k} n_j^{c'} \ln(n_j^{c'}/n)\right)}} \qquad (۷)$$

در این رابطه k تعداد جامعه، n تعداد گره‌ها، n_{ij} تعداد گره‌ها در جامعه i در جامعه‌بندی بهینه است که در جامعه‌بندی پیشنهادی در جامعه j قرار گرفته است. n_i^c تعداد گره‌ها در جامعه i در جامعه‌بندی بهینه است. $n_j^{c'}$ تعداد گره‌ها در جامعه j در جامعه‌بندی پیشنهادی است.

معیار ACC میزان صحت جامعه‌بندی پیشنهادی را تعیین می‌کند. برای محاسبه این معیار همانند معیار NMI نیاز است تا جامعه‌بندی بهینه معین باشد. این معیار با استفاده از رابطه (۸) محاسبه می‌شود [۱۷].

$$ACC = \frac{\sum_{i=1}^{n} k(c_i, PM(c_i'))}{n} \qquad (۸)$$

در این معادله، n تعداد گره‌هاست و برای گره مشخص i، c_i و c_i' به ترتیب نشان دهنده جامعه گره i در جامعه‌بندی بهینه و جامعه‌بندی پیشنهادی است. $k(x, y)$ تابعی است که در صورتیکه $x=y$ باشد، یک و در غیر این‌صورت صفر است.

۱،۴. مجموعه داده‌ها

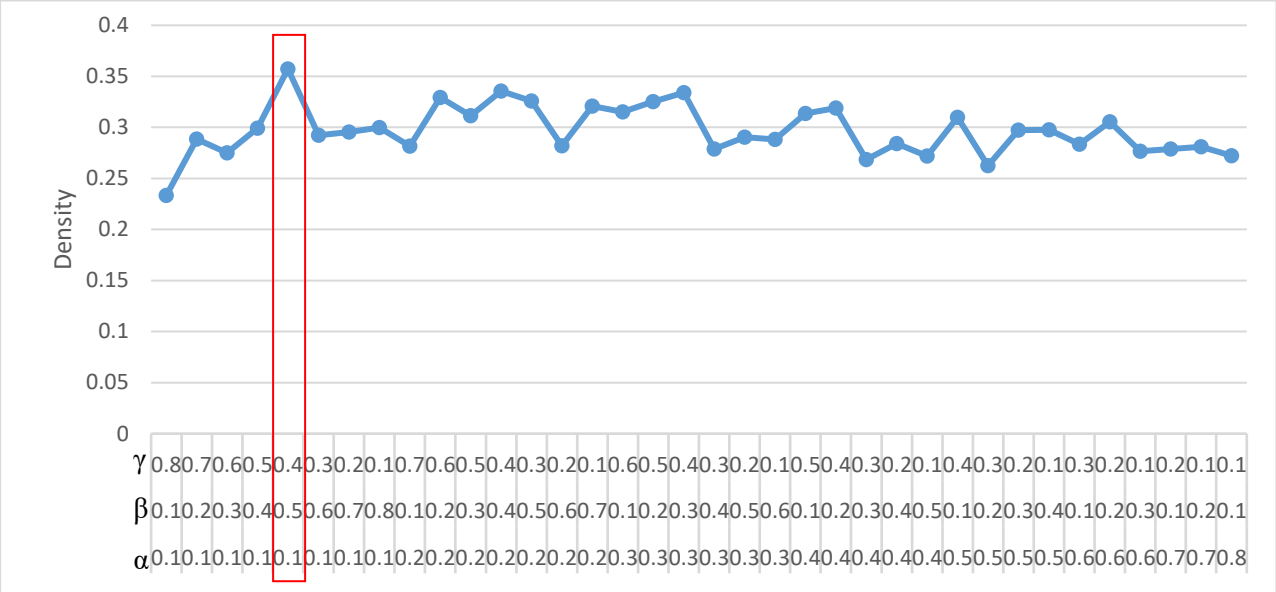
جهت بررسی کارایی روش پیشنهادی از مجموعه داده‌های شبکه اجتماعی دلفین‌ها، کالج فوتبال آمریکایی و GN استفاده شده است. شبکه دلفین‌ها، با مشاهده رفتار ۶۲ دلفین در طول هفت سال زندگی در نیوزلند و دیگر مناطق ساخته شده و حاوی ۶۲ گره و ۱۵۹ یال می باشد. این گراف به دو جامعه نر و ماده تقسیم می شود.

شبکه فوتبال حاوی ۱۱۵ گره و ۶۱۳ یال است و توسط گیروان و نیومن جمع‌آوری شده است. این داده‌ها نشان‌دهنده بازی‌های بین باشگاهی در فصل پاییز سال ۲۰۰۰ می‌باشد گره‌ها نشان‌دهنده تیم‌ها و یال‌ها نشان‌دهنده بازی بین دو تیم است. این گراف به ۱۲ جامعه گروه‌بندی شده است.

شبکهGN یک مجموعه داده مصنوعی ایجاد شده برای مسئله تشخیص جامعه، شامل ۱۲۸ گره است که در ۴ جامعه به‌طور مساوی توزیع شده است.

۲,۴. تنظیم پارامترها

پیش از آنکه به مقایسه روش پیشنهادی و روش مورد مقایسه بپردازیم، لازم است تا پارامترهای مورد استفاده در روش پیشنهادی تنظیم شود. همانطور که در قبلا اشاره شده، برای تعیین میزان تأثیر شباهت‌های SSIM، کسینوسی و همینگ از ضرایب α، β و γ استفاده شده است. برای این منظور از پارامتر چگالی و ۱۲ جامعه‌بندی مجموعه داده فوتبال استفاده شده است و مقادیر مختلف برای α، β و γ در نظر گرفته شده است که دارای شرط $\alpha+\beta+\gamma=1$ است. شکل ۵، نتیجه مقادیر متفاوت تولیدشده برای ضرایب معادله شباهت را نشان می‌دهد.



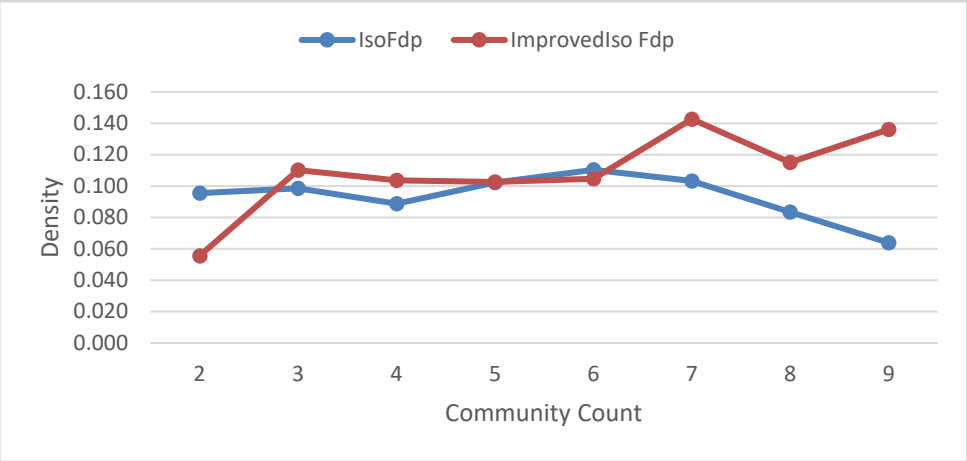
شکل ۵- بررسی ضرایب α، β و γ

همانطور که در شکل ۵ مشخص است به ازای α=0.1، β=0.5 و γ=0.4 حداکثر مقدار چگالی بدست آمده است. از اینرو در ادامه آزمایشات، مقادیر در نظر گرفته شده برای α، β و γ این مقادیر خواهد بود.

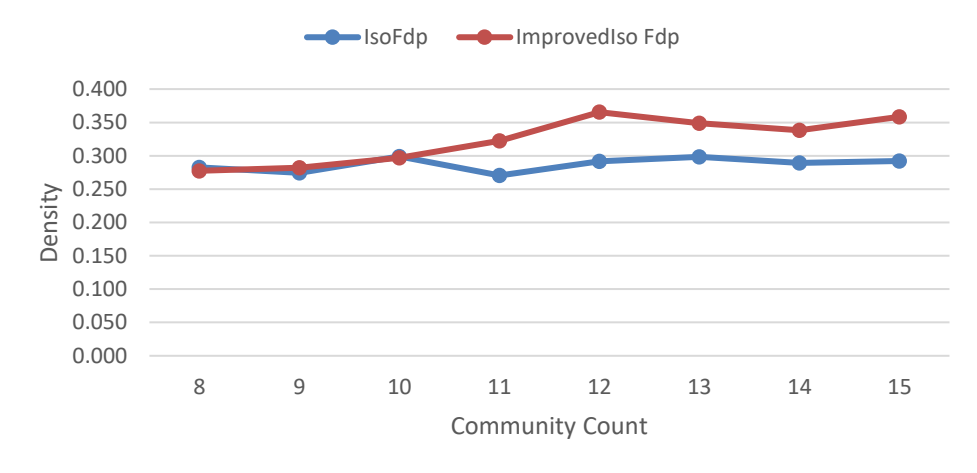
۳,۴. ارزیابی چگالی

شکل‌های ۶ تا ۸ نتایج بررسی چگالی به ازای مقادیر مختلف تعداد جامعه را به ترتیب بر روی مجموعه داده‌های دلفین، فوتبال و GN نشان می‌دهند.

نتایج شکل ۶ نشان می‌دهد که، تنها به ازای جامعه‌بندی با تعداد جوامع ۲ و ۶ عملکرد روش پیشنهادی چگالی کمتری را نشان می‌دهد اما به ازای دیگر تعداد جوامع، عملکرد روش Improved IsoFdp بهتر از روش IsoFdp است. در این شکل، با افزایش تعداد جوامع به بیش از ۶ جامعه، نمودار روش IsoFdp نزولی می‌شود که باعث پدید آمدن اختلاف میان آن و روش پیشنهادی می‌شود.



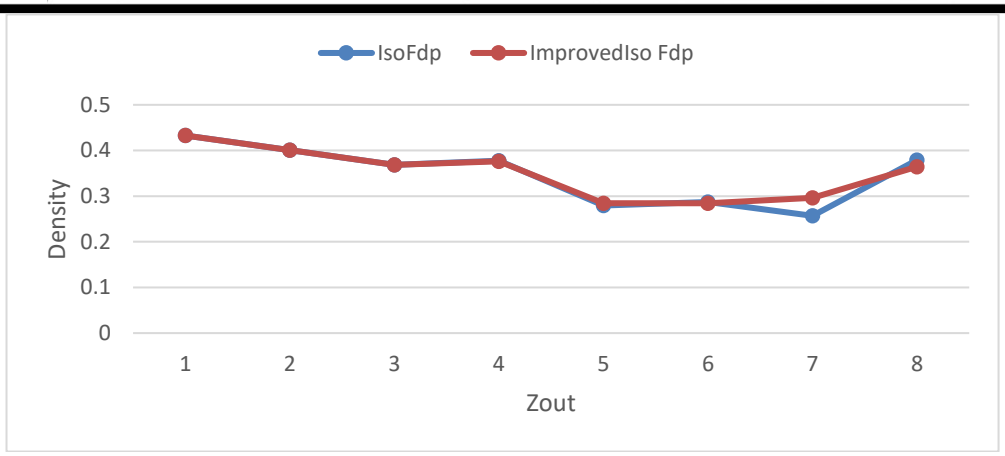
شکل ۶- چگالی به ازای تعداد جامعه مختلف در مجموعه داده دلفین



شکل ۷- چگالی به ازای تعداد جامعه مختلف در مجموعه داده فوتبال

با بررسی چگالی در مجموعه داده فوتبال، مشخص می‌شود که روش پیشنهادی در تمامی نقاط، عملکردی بهتر یا مساوی با روش مورد مقایسه دارد. از این‌رو می‌توان گفت که، در این مجموعه داده روش پیشنهادی همواره برتر از روش IsoFdp عمل می‌کند. روش پیشنهادی و روش IsoFdp در ابتدای نمودار و به ازای تعداد جوامع کمتر از ۱۱، عملکردی تقریبا یکسان را نشان می‌دهند. اما با افزایش تعداد جوامع، عملکرد روش پیشنهادی از روش پایه، فاصله می‌گیرد و تا انتهای این نمودار برتری روش پیشنهادی حفظ می‌شود (شکل ۷).

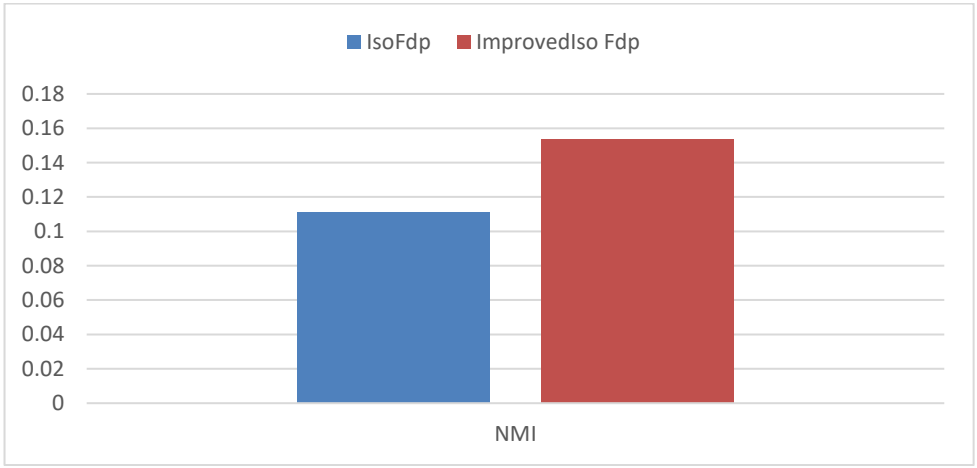
نکته قابل توجه در مورد نتایج شکل ۸ این است که، برخلاف نتایج قبلی که به ازای تعداد جوامع مختلف بیان شد، این نتایج به ازای Zoutهای مختلف است. زیرا همانطور که در بخش‌های قبلی در توضیح مجموعه داده GN بیان شد، در این مجموعه، تعداد جوامع مشخص است و تنها Zout تغییر می‌کند. نتایج این شکل نشان می‌دهد که هر دو روش تقریبا نتایج یکسانی را به ازای Zout های مختلف نشان می‌دهند. در این نمودار با افزایش Zout، چگالی روندی نزولی دارد. این موضوع، به علت پیچیده‌تر شدن عمل تشخیص جوامع، به ازای Zout های بالاتر است. اما با رسیدن به Zout 8، افزایش چگالی رخ می‌دهد که به نظر میرسد به علت به تعادل رسیدن لینک‌های ورودی/خروجی میان گره ها باشد.



شکل ۸- چگالی به ازای Zout مختلف در مجموعه داده GN

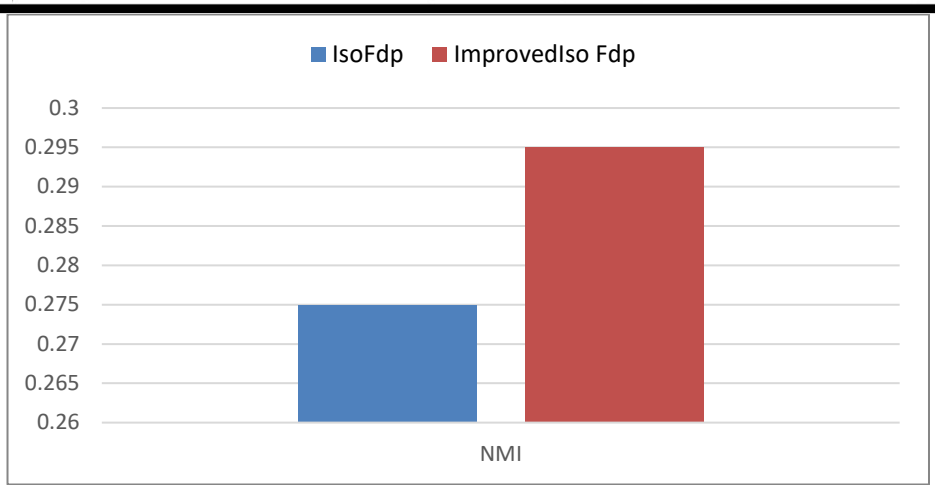
۴,۴. ارزیابی NMI

جهت محاسبه پارامتر NMI باید شبکه‌بندی بدست آمده از الگوریتم را با یک شبکه صحیح مقایسه کرد. شکل‌های ۹ تا ۱۱ حاوی اطلاعات مقایسه NMI در مجموعه داده‌های مورد بحث است. تعداد جوامع واقعی مجموعه داده دلفین، دو است. از اینرو در این آزمایش، در هر دو روش Improved IsoFdp و IsoFdp تعداد جامعه ۲ در نظر گرفته شده است تا بتوان مقدار NMI را محاسبه کرد. نتایج شکل ۹ نشان می‌دهد که NMI روش پیشنهادی، نسبت به روش مورد مقایسه بیشتر است. در شکل ۱۰ که حاصل ۱۲ جامعه مجموعه داده فوتبال است به وضوح برتری روش پیشنهادی قابل مشاهده است.

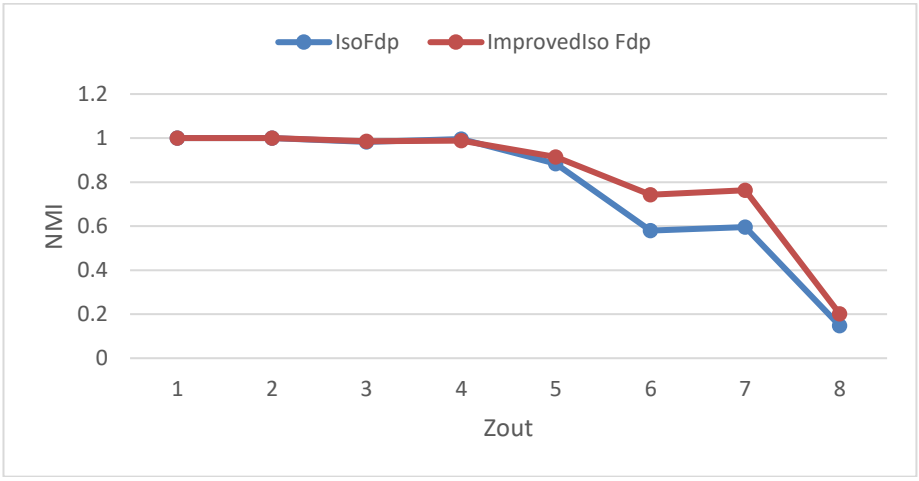


شکل ۹- NMI روش‌های مورد مقایسه در مجموعه داده دلفین

در شکل ۱۱، نتایج که به ازای مقادیر مختلف Zoutها بدست آمده است نشان می‌دهد که روش پیشنهادی به ازای مقادیر Zoutهای کمتر از ۵ نتایجی مشابه با روش IsoFdp را ارائه می‌دهد اما با افزایش Zout عملکرد روش Improved IsoFdp بهتر از روش پایه شده است. روند نزولی هر دو روش، به علت افزایش پیچیدگی جوامع، همراه با افزایش Zout است که این مسئله، تشخیص جامعه صحیح را مشکل می‌کند. اما زمانی که، Zout کم است تشخیص جامعه، به مراتب راحتتر صورت می‌پذیرد.



شکل ۱۰- NMI روش‌های مورد مقایسه در مجموعه داده فوتبال



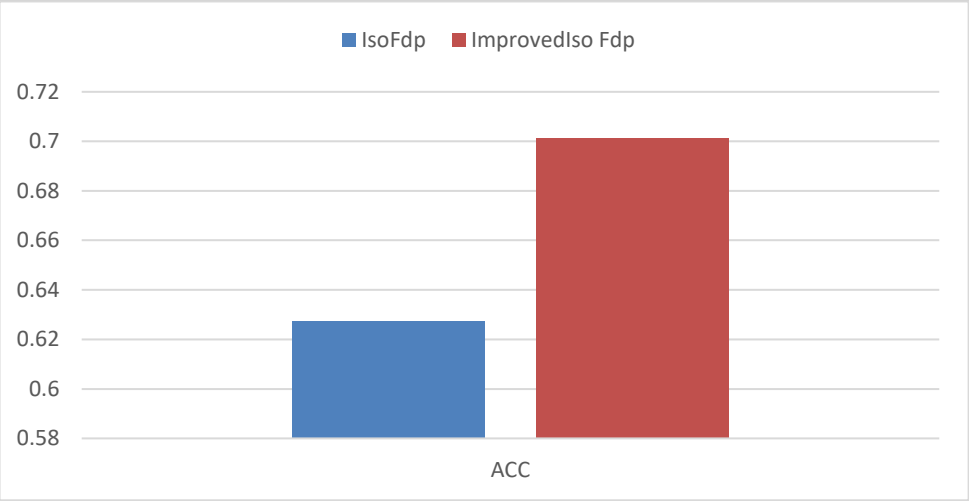
شکل ۱۱- NMI روش‌های مورد مقایسه در مجموعه داده GN به ازای مقادیر مختلف Zout

۵،۴. ارزیابی ACC

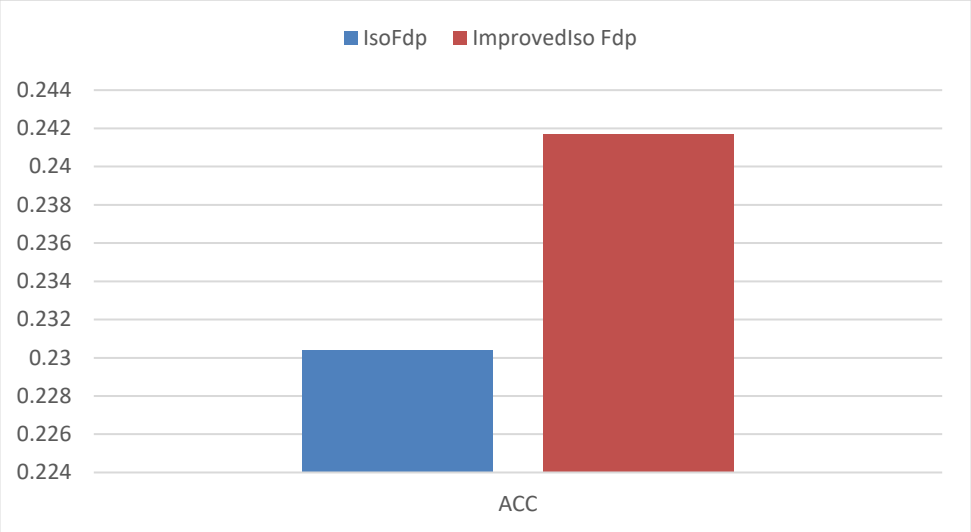
محاسبه ACC نیز همانند NMI نیازمند مشخص بودن جامعه‌بندی واقعی است. نتایج بررسی معیار ACC در شکل‌های ۱۲ تا ۱۴ ارائه شده است.

عملکرد بهتر روش پیشنهادی در شکل ۱۲ نشان می‌دهد که، صحت روش پیشنهادی بالاتر از روش مورد مقایسه در مجموعه داده دلفین بوده است. صحت روش پیشنهادی، همانند مجموعه داده دلفین، در مجموعه داده فوتبال نیز بر روش مورد مقایسه برتری دارد (شکل ۱۳).

نتایج ارائه شده در شکل ۱۴ نشان می‌دهد که تنها در $Zout = 4$ عملکرد روش پیشنهادی کمی نسبت به روش مورد مقایسه کمتر است اما به ازای مقادیر دیگر Zout، روش پیشنهادی یا با روش IsoFdp برابری و یا برتری دارد.

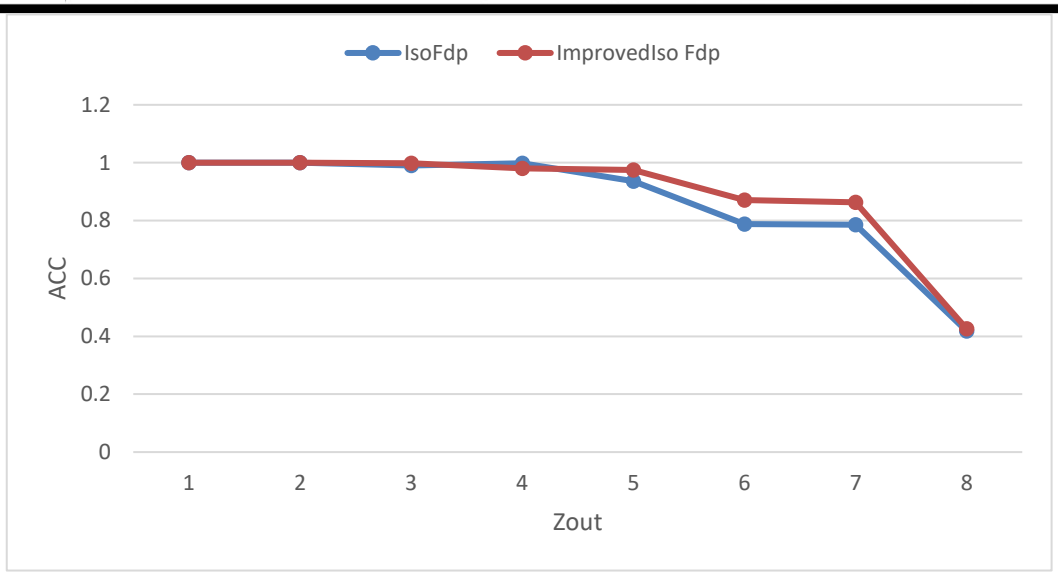


شکل ۱۲- ACC روش‌های مورد مقایسه در مجموعه داده دلفین



شکل ۱۳- ACC روش‌های مورد مقایسه در مجموعه داده فوتبال

در شکل ۱۴ نیز همانطور که انتظار می‌رفت با افزایش Zout و پیچیده‌تر شدن شبکه، دقت هر دو روش کاهش یافته است. اما نکته قابل توجه، برتری روش پیشنهادی نسبت به روش IsoFdp است.



شکل ۱۴- ACC روش‌های مورد مقایسه در مجموعه داده GN به ازای مقادیر مختلف Zout

۵. نتیجه‌گیری

در این مقاله، تلاش شده است تا روشی جهت حل مسئله جامعه‌بندی ارائه شود. روش پیشنهادی، بر اساس روش IsoFdp ارائه شده است. در روش IsoFdp از معیار شباهت ساختاری برای تشخیص میزان شباهت میان گره‌ها استفاده می‌شود. سپس از این شباهت به عنوان معیار اولیه تصمیم‌گیری برای تشخیص جوامع استفاده می‌شود.

در روش پیشنهادی تلاش می‌شود تا با بهبود معیار شباهت تأثیری منتشر شونده در تمام روش IsoFdp ایجاد شود از این‌رو در عوض استفاده از یک معیار برای مقایسه شباهت میان گره‌ها از سه معیار شباهت ساختاری، کسینوسی و همینگ استفاده می‌شود. دلیل انتخاب این سه معیار برتری آنها بر دیگر معیارهای شباهت بررسی شده در روش IsoFdp است. علاوه بر این به هر یک از این سه معیار یک ضریب اختصاص یافته است تا با تنظیم آنها وزن متناسبی به هر یک از این معیارها تعلق گیرد.

روش پیشنهادی به همراه روش IsoFdp در سه مجموعه داده دلفین، فوتبال و GN با یکدیگر مقایسه شده است. نتایج ارزیابی‌ها تحت سه پارامتر چگالی، NMI و ACC ارائه شده است. این نتایج برتری روش پیشنهادی بر روش IsoFdp را گواهی می‌دهند.

۶. مراجع

1.L, Ferreira. and L, Zhao. (2016), “Time series clustering via community detection in networks,” Information Sciences, **326**, pp 227-242.

2. C, Aggrawal. (2011),“An introduction to social network data analytics,” Social Network Data Analytics, pp 1-15.

3. S, Fortunato. and D, Hric. (2016),“Community detection in networks: A user guide,” Physics Reports, **659**, pp 1-44.

4. A, Lancichinetti. and S, Fortunato. and F, Radicchi. (2008),“Benchmark graphs for testing community detection algorithms,” PHYSICAL REVIEW E, **78**.

5. F, Malliaros. and M, Vazirgiannis. (2013),“Clustering and community detection in directed networks: A survey,” Physics Reports, **533**, pp 95-142.

6. S, Fortunato. (2010),“Community detection in graphs ,” Physics Reports, **486**, pp 75-147.

7. M, Planti_e. and M, Crampes. (2013),“Survey on Social Community Detection,” Social Media Retrieval, pp 65-85.

8. S, Fortunato. (2010),“Community detection in graphs ,” Physics Reports, **486**, pp 75-147.

9. Z, Zhao. and S, Feng. and Q, Wang. and J.Z, Huang. and G.J, Williams. and J, Fan. (2012),“Topic oriented communitise detection through social objects and link analysis in social networks,” Knowladge-Based Systems, **26**, pp 164-173.

10. M, Grivan. and M.E, Newman. (2002).“Community structure in social and biological networks,” Proceedings of the Natinal Academy of Sciences, **99**, pp7821-7826.

11. P, Zhao. and C.Q, Zhang. (2011), “A new clustring method and its application in social networks,” Pattern Recognition Letters, **32**, pp 2109-2118.

12. P, De Meo. and E, Ferrara. and G, Fiumara. and A, Provetti. (2012), “enhancing community detection using a network weighting strategy,” Information Siences.

13. A, Rodriguez. and A, Laio. (2014),“Clustering by fast search and find of density peaks,” Science, **344**, pp 1492–1496.

14. Y, Tao. and et al, (2016),“Community detection in complex networks using density-based clustering algorithm and manifold learning,” Physica A: Statistical Mechanics and its Applications, **464**, pp 221-230.

15. O, Guédon. and R, Vershynin. (2016), “Community detection in sparse networks via Grothendieck’s inequalityProbability,” Theory and Related Fields, **165**, pp 1025–1049.

16. Y, Ahn. and J.P, Bagrow. and S, Lehmann. (2010), “Link communities reveal multiscale complexity in networks,” Nature, **466** (7307), pp 761–764.

17. A, Strehl. and J, Ghosh. (2002),“Cluster ensembles a knowledge reuse framework for combining multiple partitions”, J. Mach. Learn. Res. **3**, pp 583–617.

[1] Dendogram